DOCUMENT RESUME

ED 069 783 TM 002 268

AUTHOR	Lord, Frederic M.
TITLE	Individualized Testing and Item Characteristic Curve Theory.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-72-50
PUB DATE	Nov 72
NOTE	39p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Achievement Tests; Bibliographic Citations; Bulletins; Evaluation Techniques; *Individual Tests; *Mathematical Models; *Predictive Ability (Testing); Pretesting; Probability Theory; Psychometrics; Research; Scoring Formulas; Statistics; Test Construction; Testing; *Test Validity
IDENTIFIERS	*Item Characteristic Curve Theory

ABSTRACT

An elementary survey of item characteristic curve theory, centered around the problems of individualized (tailored) testing, is presented. Following the introduction, discussions are provided of the following: Test Theory for Itemized Tests; The Guttman Scale; Item Characteristic Curve Theory; An Alternative Model; Specialization, Application, and Evaluation; Pretesting; The Statistical Estimation of Ability; A Simpler Procedure for Estimating Ability; Stochastic Approximation; The Staircase Method for Selecting the Test Questions; Scoring the Answers; Evaluation of Testing Methods; and Relation to Psychophysical Methods. An extensive list of references is provided. (DB)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 067082

RB-72-50

RESEARCH BULLETIN

INDIVIDUALIZED TESTING AND ITEM CHARACTERISTIC CURVE THEORY

Frederic M. Lord

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
November 1972

Individualized Testing and Item Characteristic Curve Theory

Abstract

An elementary survey of item characteristic curve theory is presented, centered around the problems of individualized ("tailored") testing.

Individualized Testing and Item Characteristic Curve Theory*

1. Introduction

In conventional mental testing situations, a group of individuals take the same test. Inevitably, an aptitude or achievement test is too easy for some individuals and too hard for others. Some may obtain perfect or near perfect scores on the test; others may score near zero. If successful random guessing is possible, low scores will be at and below the chance level.

If a test is too easy for some individuals, it will not discriminate effectively among them. A helpful analogy for this situation is a high-jump contest: one would not try to rank the best jumpers by always setting the bar at a level appropriate for mediocre jumpers. Similarly, if a test is too hard for some individuals, it will not discriminate effectively among them either. One would not try to rank poor jumpers by setting the bar at a level where none of them clear it.

If successful random guessing is possible (as on almost all objective tests), it is also obvious that the test cannot effectively measure an individual who gives random answers to almost all the test questions. The "noise" on his answer sheet overwhelms the "signal." This discussion suggests that for each individual there is an optimal difficulty level at which test questions are most effective for evaluating his performance or "ability."

Let us limit further consideration to the common case where all responses to test questions are (treated as) either "right" or "wrong."

---

*Preparation of this chapter was supported in part by Grant GB-32781X from the National Science Foundation.

If there is no guessing, a common rule for effectively measuring performance (this is also the rule to which theory will lead us) calls for a difficulty level such that the individual will answer half the questions correctly and half incorrectly. If questions can be answered correctly by blind guessing, then the optimal difficulty level will be somewhat easier than this.

Clearly it would be desirable to test each individual with questions best suited to his ability level. This is likely to be impractical in ordinary paper-and-pencil testing situations (but see Lord, 1971a, b, c). Now that many educational institutions have high-speed computers, however, it is becoming practical to have the computer "tailor" the test for each individual tested, administering only test questions that seem appropriate for his level of ability, as judged from his responses to the questions previously administered.

In order to tailor the test to the individual tested, the computer must be able

1. To predict from the individual's previous responses how he would respond to various questions not yet administered (these may be more, or less, difficult than any of the questions already administered).
2. To make effective use of this knowledge in picking the question to be administered next.
3. To assign at the end of the testing a numerical score (or interval estimate) somehow representing the "ability" or overall level of performance of the individual tested.

2. Test Theory for Itemized Tests

Classical test theory does not provide an appropriate framework for dealing with any of these three tasks that the computer (or the tailored-test designer) must carry out. Classical test theory is of great practical value in the design, construction, pretesting, scoring, statistical analysis, and interpretation of conventional tests of all kinds. An effective theory for similar purposes is urgently needed for individualized testing. Without careful design and appropriate scoring, individualized testing will often be inferior to conventional testing.

If we are to think meaningfully about "good" testing procedures and "inferior" procedures, we first need to be clear about the purpose of testing. The immediate purpose is not simply to determine the individual's actual performance on the particular test questions administered. This statement becomes obvious in individualized testing, since here each individual is responding to a different set of test questions, so that no comparisons among individuals are possible in terms of actual performance. Rather, the purpose is to make some inference as to his typical or expected performance on a large class of questions like those administered. In order to have a convenient label, this typical or expected performance will be called the ability of the individual in the area represented by the class of test questions.

If the questions in a class are too heterogeneous, "ability" as defined above has little psychological meaning. Science and understanding will best be served if we choose to work (at least initially) with classes

of questions sufficiently homogeneous so that we are happy to describe performance on any one class by a single number, rather than by several. This grouping of questions into homogeneous classes will be assumed in all that follows (however, see Mulaik, 1972, for a model that avoids this assumption). The reader may think of certain spelling tests, vocabulary tests, or tests of spatial abilities, among others, as providing good practical examples of homogeneous grouping of questions.

Note: There is no suggestion that an ability as defined here is in any sense a genetic, anatomical, neurological, or even psychological entity. For example, an "ability" useful in one set of circumstances as a dimension for describing individuals might in other circumstances be shown to be a composite of several abilities.

Our main problem is to infer the individual's ability (in the area represented by the test) from his performance on certain test questions. In order to do this, it is indispensable to have some idea of how the individual's responses depend upon his ability.

3. The Guttman Scale

A simple and appealing model has often been used in the attempt to describe the dependence of examinee response on examinee ability. The test questions are visualized as hurdles, the height of the hurdle being directly related to the difficulty of the question. The ability of the examinee completely determines which hurdles he can clear and which he cannot. In this deterministic model, all questions below a certain difficulty level are answered correctly by a given examinee; all questions

above this level are answered incorrectly. A scale of test questions displaying this property for all examinees is called a Guttman scale (see Torgerson, 1958, Chapt. 12).

This model is arrived at by asking what we would like a test to do. It would be nice if we could know from the examinee's test score (number of right answers) exactly how he responded to every question in the test. This knowledge can be obtained from a Guttman scale but not from any other kind of test.

Although approximate Guttman scales are of use in sociological work and in attitude measurement, they seem to be of little interest in aptitude and achievement testing. For one thing, in many common situations an ideal aptitude or achievement test should have all items of equal difficulty. According to the deterministic hurdle model, all examinees should obtain either a zero score or a perfect score on such a test. Nothing like this happens in practice, however. The distribution of number-right scores is typically bell-shaped, even when we try by every means to obtain a U-shaped distribution.

The Guttman scale assumes that the tetrachoric correlation between scores on any two test questions is 1.00. For two questions of medium difficulty, this would mean a product-moment correlation of approximately 1.00 also. Actually, the tetrachoric correlation between typical aptitude or achievement test questions is not 1.00 but only about .15. The product-moment correlation between questions of medium difficulty is about .10.

4. Item Characteristic Curve Theory

If we want a mathematical model capable of fitting typical aptitude or achievement test data, we must use a probabilistic rather than a deterministic model. Denote the probability that individual a will answer a test question correctly by $P_{\underset{\sim}{\beta}}(\theta_a) \equiv \text{Prob}(U_a = 1 | \theta_a, \underset{\sim}{\beta})$. Here U_a is a random variable that assumes the value 1 when individual a answers correctly, 0 otherwise. The real number θ_a represents the ability of individual a. The vector $\underset{\sim}{\beta}$ contains parameters fully characterizing the test question ("item") administered. The difficulty of the item, for example, will be represented by one of the parameters in $\underset{\sim}{\beta}$. All this notation serves only to assert that the probability that individual a will answer a question correctly depends only upon the ability of the individual and upon certain characteristics of the test question.

The probability $P_{\underset{\sim}{\beta}_i}(\theta_a)$ is to be interpreted here (see next section) as a relative frequency over randomly selected test questions all having the same characteristics $\underset{\sim}{\beta} = \underset{\sim}{\beta}_i$. There is no consideration here of repeated testing--each individual is tested only once.

It is natural to assume that $P_{\underset{\sim}{\beta}}(\theta)$ is an increasing function of θ. The higher the ability level, the greater the probability of a correct answer. This will be assumed hereafter. Some typical functions $P_{\underset{\sim}{\beta}}(\theta)$ (see Lord, 1968) are shown in Figure 1 for illustrative purposes.

We wish to assume that the probability of a correct answer to a question depends <u>only</u> on the individual's ability level and on $\underset{\sim}{\beta}$, not on any other known characteristic of his, nor on any other characteristics of

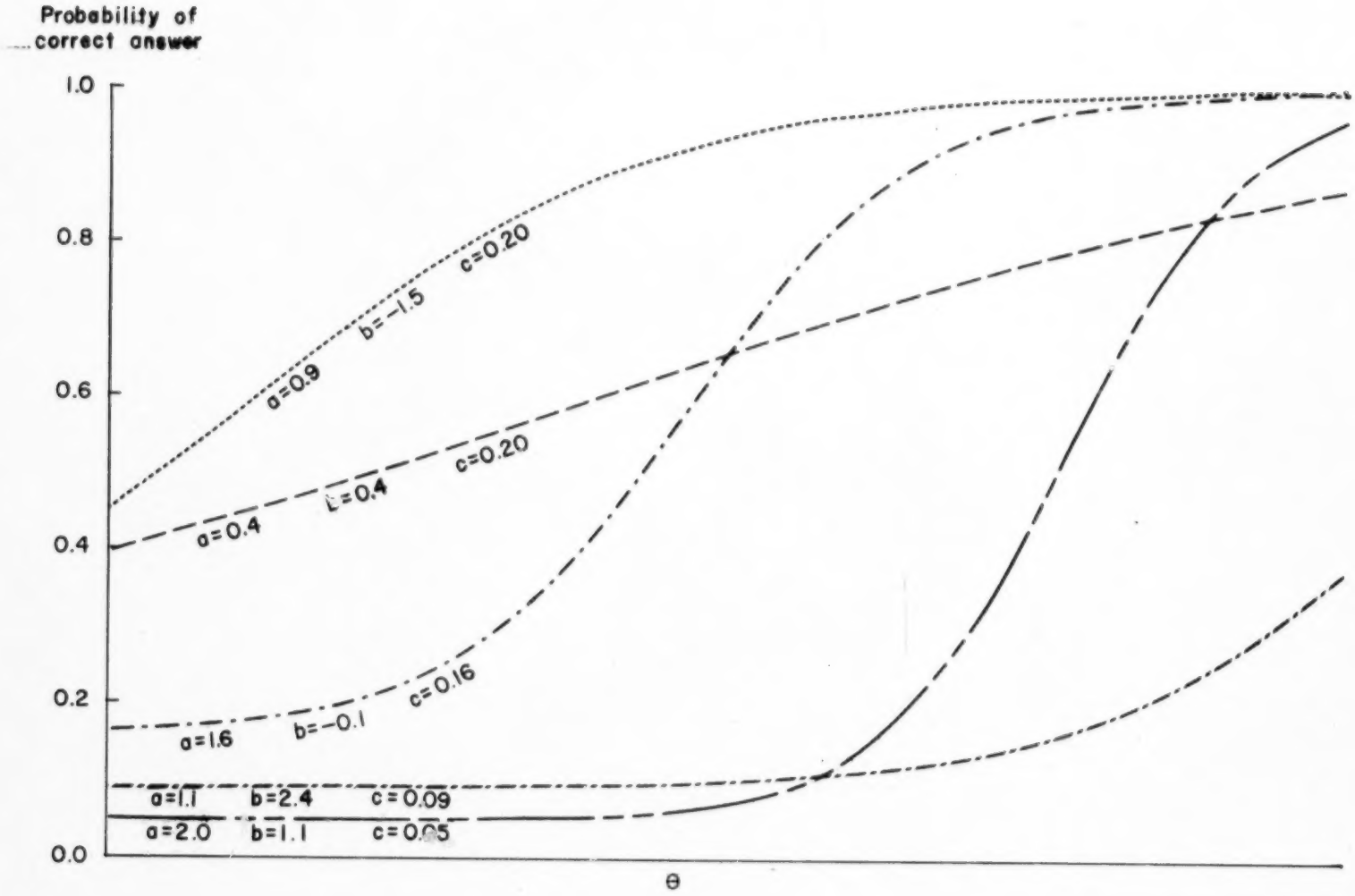

Figure 1. Probability of correct answer as a function of ability, as estimated for five SAT verbal questions.

the question, nor on any other variable available to us. It follows that the probability of an individual answering correctly is not altered by knowledge of the actual performance of other individuals. Thus, for example, the probability of correct answers to a question by both individuals a and a' is given by the product $P_{\underset{\sim}{\beta}}(\theta_a)P_{\underset{\sim}{\beta}}(\theta_{a'})$.

In addition, it follows that the probability of an individual answering a question correctly is not altered by knowledge of his actual performance on other questions. Thus, for example, the probability that individual a will answer questions i , i' , and i" all correctly is given by the product $P_{\underset{\sim}{\beta}}(\theta_a)P_{\underset{\sim}{\beta}'}(\theta_a)P_{\underset{\sim}{\beta}''}(\theta_a)$. This is called the principle of local independence (Lazarsfeld, 1959).

It is instructive to see what would happen if local independence did not hold. Suppose that for a certain individual a the probability that he will answer randomly chosen questions i , i' , and i" all correctly is greater than $P_{\underset{\sim}{\beta}}(\theta_a)P_{\underset{\sim}{\beta}'}(\theta_a)P_{\underset{\sim}{\beta}''}(\theta_a)$. If this is not a unique occurrence, this would mean that there are individuals at ability level $\theta = \theta_a$ who score systematically higher on these test questions than other individuals with the same θ level. Thus these test questions would be measuring some psychological dimension other than θ . This is just the situation that the assumption of local independence is designed to exclude. We want to deal with a test that measures the ability θ ; we do not want to deal (at least at first) with a test score that may represent either of two (or more) psychological dimensions at once.

5. An Alternative Model

It seems necessary at this point to mention another model very commonly confused with the $P_{\beta}(\Theta_a)$ model used here. This other model makes assertions about the probability, to be denoted by $P_i(\Theta_a)$, that a specific individual a answers a specific item i correctly. The two models will be distinguished and reasons given for discarding one of them.

If an individual responds to question i at random, it is clear that his probability of success is the reciprocal of the number of possible responses to question i. There are many questions, however, for which this individual knows the correct answer; for such a question, his probability of answering correctly would seem to be virtually 1. There may be other questions on which this individual is misinformed; for such a question, his probability of answering correctly would seem to be virtually 0.

Consider two individuals, a and b, and two test questions, i and j. Individual a happens to know the answer to question i and to be misinformed on question j. Individual b happens to know the answer to question j and to be misinformed on question i. If we write $P_i(\Theta_a)$ for the probability that individual a answers question i correctly we have $P_i(\Theta_a) = 1$, $P_j(\Theta_a) = 0$, $P_i(\Theta_b) = 0$, $P_j(\Theta_b) = 1$, approximately. The first two equations considered together imply that question i is easier than question j, the last two equations imply just the reverse. Thus questions i and j must measure a different ability for individual a than they measure for individual b. This is a possible model (Meredith, 1965) and a possible interpretation, but usually not a fruitful one, since usually we want to compare individuals a and b along the same ability dimension.

In order to avoid the situation just outlined, we will use only the model defined in the previous section, which makes no assertions about the probability $P_i(\theta_a)$ that a specified individual a answers a specified test question i correctly. The model that we will use deals instead with $P_{\underset{\sim}{\beta}}(\theta_a)$, which represents the long-run relative frequency of correct answers given by individual a when answering test questions all having the same specified $\underset{\sim}{\beta}$. An equivalent statement is that $P_{\underset{\sim}{\beta}}(\theta_a)$ represents the probability that individual a will answer correctly a question chosen at random from all questions having the same $\underset{\sim}{\beta}$. When the model holds, the function $P_{\underset{\sim}{\beta}}(\theta)$ of θ will be referred to as the characteristic curve for each item having parameters $\underset{\sim}{\beta}$.

6. Specialization, Application, and Evaluation

Empirical checks on the validity and practical utility of the item characteristic curve (icc) model have in large part been delayed for about twenty-five years because of the difficulty of estimating the characteristic curves of particular items. Recently a number of workers have successfully estimated many icc and some evidence of the validity and usefulness of the model has been accumulated. The present section is intended to refer the reader to materials relevant for assessing the validity and usefulness of the model; no detailed discussion is possible here.

An approach that estimates icc without restrictive assumptions about their mathematical form has been described by Lord (1970a). If it can be assumed simply that the icc differ only by a linear transformation of θ (a common assumption), a computer program implementing Levine (1972) has been found very effective for estimating icc (Levine, personal communication).

It has been common to assume that the icc are normal ogives or logistic curves. If the icc are logistic and if, for a given test, the curves all have the same slope parameter, the present model can be shown (Birnbaum, 1968, p. 402) to be the same as the well-known Rasch model, which has certain desirable measurement properties (Rasch, 1960, 1961, 1966a, b; Wright, 1968; Wright and Panchapakesan, 1969). Methods for estimating the single item parameter needed in this model and studies evaluating the fit and effectiveness of this model have been reported by Rasch, by Wright and Panchapakesan, and by Lawley (1943, 1944), Andersen (1970, 1971a, b, 1972a, b), Anderson, Kearney, and Everett (1968), Choppin (1968), Fischer (1972), Fischer and Scheiblechner (1970), Hambleton (1969), Hambleton and Traub (1971), Panchapakesan (1969), Scheiblechner (1971a, b), Tinsley and Dawis (1972), Urry (1970). Reports on the fit and effectiveness of the one-parameter model range from disapproval to enthusiasm.

If some test questions correlate higher with ability than others, as is commonly the case, a one-parameter model may be inadequate. Whenever correct answers can be obtained by random guessing, even a two-parameter model is likely to be inadequate. Modified normal ogive and logistic models with three parameters are available (Birnbaum, 1968, chapter 17). The mathematical formulas are

$$P_{\underset{\sim}{\beta}}(\theta) = \gamma + (1 - \gamma) \int_{-\infty}^{\alpha(\theta-\beta)} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2} t^2)\, dt \qquad (1)$$

for the modified normal ogive, and

$$P_{\underset{\sim}{\beta}}(\theta) = \gamma + \frac{1 - \gamma}{1 + \exp[-1.7\alpha(\theta - \beta)]} \qquad (2)$$

for the modified logistic. Models (1) and (2) do not differ anywhere by more than .01. We need not debate here which model is more nearly correct. Neither, of course, is exactly correct.

The three parameters in $\underset{\sim}{\beta} \equiv \{\alpha, \beta, \gamma\}$ may be thought of as

α discriminating power (a measure of the relation between item score and ability),

β difficulty,

γ probability of a correct answer for individuals at lowest ability levels.

A more detailed, practical discussion of these item parameters is given in Lord (1970b).

Lord and Novick (1968, section 16.11) consider for what practical situations the normal ogive model is likely to be appropriate. Studies evaluating the fit of the model to actual test data include Lord (1952, 1970a, 1972); Indow and Samejima (1962, 1966). Their findings support the model for the data studied. Many more evaluative studies are needed.

Methods for estimating the item parameters have been developed and tried out by Lord (1952, 1968, 1972), Indow and Samejima (1962, 1966), Birnbaum (1968), Bock (1970, 1972), Bock and Lieberman (1970), Kolakowski and Bock (1970), Kolakowski (1969, 1972), Lees, Wingersky, and Lord (1972). Studies making theoretical or practical use of these models appear in two books by Solomon (1961, 1965). Included among other such studies are those by Brogden (1946), Tucker (1946), Cronbach and Warrington (1952), Lord (1953a, b; 1955, 1970a, b; 1971a, b, c, d, e), Cronbach and Merwin (1955), Anderson (1959), Merwin

(1959), Cronbach and Azuma (1962), Paterson (1962), Birnbaum (1968, chapters 17-20; 1969), Wood and Skurnik (1969), Shiba (1969a, b), Shoemaker and Osburn (1970), Urry (1970), Nishisato and Torii (1971), Bay (1971). Hambleton and Traub (1971), Samejima (1972).

7. Pretesting

In order to design a test for the specific purpose of measuring the ability of a particular individual, we must have available a large pool of test questions that have been extensively pretested, so that the parameters characterizing each question may be considered known. Note that the item characteristic function $P_{\underset{\sim}{\beta}}(\theta)$ does not depend on the distribution of ability in any group of individuals. Consequently, the parameters $\underset{\sim}{\beta}$ for a test question can be determined once and for all by pretesting in some convenient group. Of course, reliance on the robustness of the model over wide variations in group should not be carried to extremes. In practice, the pretest group should resemble the collection of individuals who will later be given the individualized tests.

Corresponding to the invariance of the item parameters $\underset{\sim}{\beta}$ over groups of individuals there is an invariance of the ability parameter θ over different tests (cf. Rasch, 1961, pp. 331-333). These invariances are fundamental to the success of item characteristic curve theory in comparison with older item analysis methods. In a leading older method, each item would be characterized by the proportion of correct answers received and by the correlation between item response and total test score. However, these item parameters of the older method would be different for different pretest groups; also, the correlation parameter would

change if the test was lengthened or otherwise modified. This lack of invariance limits the usefulness of classical item analysis. The usual kinds of test score for an individual have a similar lack of invariance when the test administered is modified.

8. The Statistical Estimation of Ability

Once the item parameters $\underset{\sim}{\beta}$ have been determined by pretesting, the problem of estimating the ability of an individual from his responses is a straightforward statistical estimation problem. If his probability of success on question i is $P_{\underset{\sim}{\beta}_i}(\theta)$ and his probability of failure is $Q_{\underset{\sim}{\beta}_i}(\theta) \equiv 1 - P_{\underset{\sim}{\beta}_i}(\theta)$, then the likelihood function for his score ($u_i = 1$ or 0) on question i is simply

$$L_i(\theta) = \begin{cases} P_{\underset{\sim}{\beta}_i}(\theta) & \text{if } u_i = 1 \\ Q_{\underset{\sim}{\beta}_i}(\theta) & \text{if } u_i = 0 \end{cases} .$$

This may be more conveniently written

$$L_i(\theta) = [P_{\underset{\sim}{\beta}_i}(\theta)]^{u_i}[Q_{\underset{\sim}{\beta}_i}(\theta)]^{1-u_i} .$$

Because of local independence, the likelihood function for the individual's responses to a test of n questions is simply the product of the likelihoods for the separate questions:

$$L(\theta) = \prod_{i=1}^{n} [P_{\underset{\sim}{\beta}_i}(\theta)]^{u_i}[Q_{\underset{\sim}{\beta}_i}(\theta)]^{1-u_i} .$$

Since the $\underset{\sim}{\beta}_i$ are known from pretesting, it is not difficult for a computer, given a mathematical form for $P_{\underset{\sim}{\beta}}(\theta)$ such as (1) or (2), to find the maximum likelihood estimate $\hat{\theta}$ of the individual's ability. ($\hat{\theta}$ is the value of θ that maximizes the likelihood $L(\theta)$ of his observed responses $u_1, u_2, \cdots, u_n$.)

9. A Simpler Procedure for Estimating Ability

There still remains the problem of how to pick the n test questions to be administered to a given individual. One advantage of individualized testing is that testing can be continued until the individual's ability has been estimated with some predetermined degree of statistical accuracy. For the sake of simplicity, however, we will consider here only the case where n is fixed.

To make matters even more simple, let us select from the pool a large set of pretested questions that differ from each other only in difficulty (β). These questions have identical values of α and γ . If (1) or (2) held with no random guessing, the optimal test for estimating θ with minimum squared error would consist entirely of questions for which $P_{\underset{\sim}{\beta}}(\theta_a) = \frac{1}{2}$, where θ_a is the ability of the individual to be tested. Since we do not know θ_a and cannot estimate it with any accuracy in advance of testing, all this would not give us a method for choosing the n test questions to be administered. Such methods will be discussed in the next section.

Let us assume now (as seems reasonable) that individual ability (θ) and item difficulty (β) are measured along the same dimension, in the sense that for any increment k an increase in ability from θ to $\theta + k$ could hypothetically be exactly offset by an equivalent increase in difficulty from β to $\equiv \beta + k$. In other words, $P_{\{\alpha, \beta, \gamma\}}(\theta)$ and $P_{\{\alpha, \beta+k, \gamma\}}(\theta + k)$ represent exactly the same function of θ . This assumption holds for models (1) and (2) and for any other $P_{\underset{\sim}{\beta}}(\theta)$ in which θ and β appear only as their difference $\theta - \beta$. Under this assumption $P_{\underset{\sim}{\beta}}(\theta) \equiv F(\theta - \beta)$ where F is an unspecified monotonic function.

What we have assumed here is simply that we have a large set of questions, selected from the pretested pool, whose icc differ only by a translation along the θ axis. Let us define β^o as the item difficulty level at which the individual has probability of success $F(0)$. Thus $\theta = \beta^o$. We can determine an individual's ability θ by determining his β^o .

It is possible in practice to find the proportion of correct answers actually given by individual a to test questions at any specified difficulty level. By trial and error, or by better methods to be discussed below, we can in this way find approximately the difficulty level β_a^o such that $P_{\{\alpha, \beta_a^o, \gamma\}}(\theta_a) = F(0)$. This difficulty level <u>is</u> (approximately) the ability level of individual a , since by definition $\theta_a = \beta_a^o$.

10. Stochastic Approximation

Clearly what we need now is some method better than trial and error for finding β^o . Stated in this way, the problem is a standard problem

in stochastic approximation (Wasan, 1969). Specifically, the stochastic approximation problem is to select a sequence of test questions so that we can conveniently and accurately estimate the individual's ability θ_a from the sequence $u_1, u_2, \ldots, u_n$ of responses. Since for simplicity we have selected from the pretest pool a set of test questions that differ statistically only on their difficulty parameter (for a treatment that avoids this, see Owen, 1970), the problem of selecting a sequence of test questions is simply the problem of selecting a sequence $\beta_1, \beta_2, \ldots, \beta_n$. The resulting sequence of questions constitutes an individualized test or tailored test designed for effective measurement of the particular individual tested.

The difficulty β_1 of the first question administered can be chosen in the same way that we would choose the average difficulty level of the questions in a conventional test--by subjective judgment or by using a Bayesian prior. If the individual answers the first question incorrectly, we guess that it is too hard for him and choose an easier question to administer next. If he answers the first question correctly, we guess that it is too easy for him and choose a harder question to administer next.

After administering the second question, we could use the statistical method outlined in section 8 to obtain from his responses to the first two questions an estimate $\hat{\theta}_a^{(2)}$ of the individual's ability. The difficulty of the third question administered could be matched to the individual's estimated ability by choosing $\beta_3 = \hat{\theta}_a^{(2)}$. We could then choose $\beta_4, \beta_5, \ldots$ similarly. However, a procedure that proceeds by steps that are individually optimal is not in this case likely to be an optimal procedure overall.

We will not try here to devise an optimal procedure. Rather, we will try to find a good, simple procedure that is not only easy to carry out but also easy to evaluate as a procedure for statistical inference.

Under the Robbins-Monro stochastic approximation procedure, the rule for choosing the difficulty of the $(v + 1)$ -st question is

$$\beta_{v+1} = \beta_v + d_v(u_v - F(0)) \quad , \tag{3}$$

where $d_1, d_2 \ldots$ is a suitable decreasing sequence of positive numbers chosen in advance (Robbins and Monro, 1951). If the step size d_v is small, the $(v + 1)$ -st question will be chosen to have nearly the same difficulty as the v -th question; if d_v is large, there will be a more substantial change in difficulty. In the Robbins-Monro procedure, the d_v are chosen relatively large initially when little is known about the individual's ability level, allowing substantial readjustments in item difficulty levels. Later when the appropriate difficulty level has been approximated, the chosen d_v are small, eventually approaching zero. Typically $d_v = d_1/v$, $v = 1,2,3,\ldots$.

Robbins and Monro's proof shows that when (3) is used with suitable d_v , the item difficulty β_{v+1} is a consistent estimator of the individual's ability θ , in the sense that β_{v+1} converges stochastically to θ as v becomes large. Formulas leading in some cases to asymptotically optimal choices of the d_v are given by Hodges and Lehmann (1956).

11. The Staircase Method for Selecting the Test Questions

Unfortunately, the Robbins-Monro procedure requires storing 2^n test questions in the computer before testing is begun, where n is the

number (here assumed to be fixed in advance) of questions to be administered to the examinee. For most aptitude and achievement tests composed of dichotomously scored questions, $n \geq 25$.

An alternative procedure, keeping the total number of test questions within acceptable limits, is available: the up-and-down method or staircase method, used in testing explosives, in bioassay, in psychophysics, and elsewhere. In the up-and-down method, the rule for selecting questions is still given by (3), but with d_v replaced by a constant <u>step size</u> d .

If $F(0) = 1/2$, the up-and-down rule becomes

$$b_{v+1} = \begin{cases} b_v + d & \text{if question } v \text{ is answered correctly,} \\ b_v - d & \text{if question } v \text{ is answered incorrectly.} \end{cases}$$

This simple form of (3) normally holds only if there is no guessing of correct answers.

For basic discussions of this method, see Dixon and Mood (1943) and Brownlee, Hodges and Rosenblatt (1953). Some modifications are discussed by Tsutakawa (1963, 1967a, b).

This method requires storing only $n(n + 1)/2$ test questions in the computer in advance of testing. This number can be reduced further by taking a few obvious shortcuts.

12. Scoring the Answers*

Consider the following three simple methods for scoring the student's responses to the test questions:

---

*This section and part of the previous section are a slight revision of material appearing in Lord (1971e).

1. The "final-difficulty score," β_{n+1}, the difficulty of the $(n+1)$-th question (not actually administered) as defined by equation (3).

2. The "number-right score," $\sum_{v=1}^{n} u_v$, or the "proportion-right score," $\frac{1}{n}\sum_{v=1}^{n} u_v$. The former is the score most commonly used in scoring conventional mental tests.

3. The "average-difficulty score," $\bar{\beta} = \frac{1}{n}\sum_{v=2}^{n+1} \beta_v$. This score is simply the average of the difficulty parameters of the questions administered, omitting the first (since the first question is the same for all individuals tested) and including β_{n+1}.

[Before going ahead, the reader may wish to make a guess as to the relative merits of these three scoring methods for the up-and-down (fixed step size) procedure.]

When the step size shrinks appropriately as n increases, as in the Robbins-Monro procedure, β_{n+1} is a good estimator of ability. When the step size is fixed, as in the up-and-down method, β_{n+1} is no longer a consistent estimator for θ, nor does its sampling variance approach zero as n becomes large. It turns out that when step size is fixed, number-right score is perfectly correlated with β_{n+1}; so it, too, can be eliminated as an effective method of scoring.

Brownlee, Hodges, and Rosenblatt (1953) have shown that the average-difficulty score is asymptotically equivalent to the maximum likelihood

estimator for θ found by Dixon and Mood (1943) for the up-and-down method. Although no optimum small-sample properties have been proven for the average-difficulty score, it appears at present to be the preferable method of scoring tests administered by the up-and-down method.

It frequently happens that similar groups of students are tested year after year. In this case, an excellent prior distribution for the parameter θ is available, based on records of past performance. In such situations, the careful design of a tailored testing procedure would certainly be based on a Bayesian approach. The Bayesian approach will not be treated here since it is of greater mathematical complexity. The interested reader is referred to Owen (1970) and to Freeman (1970).

13. Evaluation of Testing Methods

The remaining problem for discussion here is the evaluation of different stochastic approximation procedures and of different choices of parameters such as d.

Properties of the Robbins-Monro procedure for large n are discussed in the references given. Some properties for small n are treated by Wasan (1969, chapt. 2) and by Cochran and Davis (1965). An improved procedure for small n is suggested by Kesten (1958) and tried out empirically by Odell (1961).

The up-and-down rule for selecting test questions to be administered produces a Markov chain or, more specifically, a random walk for the values of β_v. The transition probabilities $P_{\{\alpha,\beta_v,\gamma\}}(\theta)$ and $Q_{\{\alpha,\beta_v,\gamma\}}(\theta)$

are stationary. They depend on β_v, but they do not depend on v when β_v and θ are given.

Starting from this, it is not hard to write down a formula for the frequency distribution of $\beta_{(n+1)}$ under the up-and-down method; but $\beta_{(n+1)}$ is not a satisfactory scoring procedure for this method, as already noted. The frequency distribution of the average difficulty score $\bar{\beta}$ is not easily obtained for moderate n, but Brownlee, Hodges, and Rosenblatt have provided recursive formulas from which the mean and sampling variance of $\bar{\beta}$ can be readily calculated numerically by computer for given θ, α, β_1, γ, d, $F(0)$ and for any n likely to be of interest. Given the bias and sampling variance of $\bar{\beta}$ for given θ for each of various testing designs, it is not hard to decide which design is preferable for measuring at a specified ability level.

A variety of testing designs were investigated in this fashion by Lord (1970b, 1971d). Numerical studies of a variety of stochastic approximation methods applicable to individualized testing are reported by Cochran and Davis (1964), Davis (1971), Wetherill (1963), Wetherill and Levitt (1965), Wetherill, Chen, and Vasudeva (1966). Other empirical studies of individualized testing include Bayroff and Seeley (1967), Ferguson (1971), Hansen and Schwarz (1968), Linn, Rock, and Cleary (1969, 1972), Paterson (1962), Seeley, Morton, and Anderson (1962), Urry (1970), Waters (1964), Waters and Bayroff (1971), Wood (1969).

14. Relation to Psychophysical Methods

The up-and-down method is often used in bioassay. According to Guilford (1954), it originally was developed for the study of explosives. When used in psychophysical studies, it is known as the staircase method (Cornsweet, 1962).

The psychophysicist does not need to know the precise mathematical form of the psychometric function. To elucidate comparison with icc theory, let us assume that the psychometric function actually is given by equation (1) or (2) with $\gamma = 0$.

Whereas the mental tester controls α and β (by using pretested items) while trying to estimate the value of θ, the psychophysicist (or bioassayist) controls θ while trying to estimate the value of β and, sometimes, the value of α. Note that θ and β play reversed roles for the mental tester and for the psychophysicist. For the latter, θ might represent the physical intensity of the various stimuli presented under experimental control. Then, β would be the "threshold" at which the subject says "yes, I detect the stimulus" $F(0)$ of the time; α would be the precision of the psychometric function. The psychophysicist chooses the stimulus level θ_1, administers this stimulus, and records the response $u_{i1} = 0$ or 1. He then chooses another stimulus level θ_2, administers this stimulus, records $u_{i2} = 0$ or 1, and continues in this way.

In mental testing, we are interested only in the relative values of θ for different examinees; θ is, at best, measured on an interval scale. The unit and zero point of this scale have little ready meaning for most other scientists concerned with mental measurement. The psychophysicist,

on the contrary, usually estimates the absolute value of β on some standard scale having a unit and origin well known to physicists and other scientists.

Avoiding bias in his estimates is therefore of crucial importance for the psychophysicist. In mental testing, any linear transformation of θ is as valid as any other. Bias is usually of no importance to the mental tester as long as it affects all scores equally.

The fact that the psychophysicist has two unknown parameters, α and β, creates a further problem. It is not possible for him to choose the step size d optimally without knowing α. A poor choice of d leads either to excessive standard error or bias in the estimated threshold, or else to experiments that are unnecessarily lengthy.

Often the psychophysicist can obtain observations cheaply. It may be easy for him to obtain a thousand or ten thousand responses from a single subject. The mental tester cannot do this. The objective situation forces the mental tester to use reasonably efficient testing and estimation methods. For the psychophysicist, statistically efficient procedures may be unnecessary and distinctly uneconomical.

In addition to the staircase method, the psychophysicist sometimes uses block up-and-down methods (Stuckey, Hutton, and Campbell, 1966; Tsutakawa, 1963, 1967a, b; Cochran and Davis, 1964) and unequal-step-size "sequential" methods (Taylor and Creelman, 1967; Pollack, 1968). The time-honored constant-stimulus method corresponds in part to conventional (not individualized) mental testing; the scoring methods are different in the two applications, however.

The indicated correspondence between individualized testing and certain psychophysical experiments is clear and instructive whenever the mental

tester can work with items all having the same α and γ accurately determined by pretesting. Such situations do not really exist at present, however. Not enough items are usually available to do practical work with a pool of items all having the same α and γ (proponents of Rasch's method may disagree).

Present work in icc theory and practice is concerned with estimating item and examinee parameters simultaneously. This is very different from the typical psychophysical problem. An outstanding current problem is how to carry out individualized testing using test items characterized by a variety of inaccurately estimated item parameters. A recent article by Dupač and Král (1972) is relevant for individualized testing with fallibly estimated values of β_i.

References

Andersen, E. B. Conditional inference for multiple choice questionnaires. Report No. 8. Copenhagen: Copenhagen School of Economics and Business Administration, 1970.

Andersen, E. B. A goodness of fit test for the Rasch model. Report No. 9. Copenhagen: Copenhagen School of Economics and Business Administration, 1971. (a)

Andersen, E. B. Conditional inference and models for measuring. Copenhagen: Copenhagen School of Economics and Business Administration 1971. (b)

Andersen, E. B. The numerical solution of a set of conditional estimation equations. Journal of the Royal Statistical Society, Series B, 1972, 34, 42-54. (a)

Andersen, E. B. A computer program for solving a set of conditional maximum likelihood equations arising in the Rasch model for questionnaires. Research Memorandum 72-06. Princeton, N.J.: Educational Testing Service, 1972. (b)

Anderson, J., Kearney, G. E., and Everett, A. V. An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Anderson, T. W. Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.), Probability and statistics. New York: Wiley, 1959.

Bay, K. S. An empirical investigation of the sampling distribution of the reliability coefficient estimates based on alpha and KR20 via computer simulation under various models and assumptions. Unpublished doctoral dissertation, Edmonton, Alberta, University of Alberta, 1971.

Bayroff, A. G., and Seeley, L. C. An exploratory study of branching tests. Technical Research Note 188. Washington, D.C.: U.S. Army Behavioral Science Laboratory, 1967.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Pp. 397-479.

Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. Journal of Mathematical Psychology, 1969, 6, 258-276.

Bock, R. D. Estimating multinomial response relations. In R. C. Bose (Ed.), Contributions to statistics and probability essays in memory of Samarendra Nath Roy. Chapel Hill, N.C.: University of North Carolina Press, 1970. Pp. 453-479.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Bock, R. D., and Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. Psychometrika, 1946, 11, 197-214.

Brownlee, K. A., Hodges, J. L., Jr., and Rosenblatt, M. The up-and-down method with small samples. Journal of the American Statistical Association, 1953, 48, 262-277.

Choppin, B. H. An item bank using sample-free calibration. Nature, 1968, 119, 870-872. Reprinted in R. Wood and L. S. Skurnik, Item banking. Slough: National Foundation for Educational Research in England and Wales, 1969. Pp. 134-140.

Cochran, W. G., and Davis, M. Stochastic approximation to the median effective dose in bioassay. In J. Gurland (Ed.), Stochastic models in medicine and biology. Madison: University of Wisconsin Press, 1964. Pp. 281-300.

Cochran, W. G., and Davis, M. The Robbins-Monro method for estimating the median lethal dose. Journal of the Royal Statistical Society, Series B, 1965, 27, 28-44.

Cornsweet, T. N. The staircase method in psychophysics. American Journal of Psychology, 1962, 75, 485-491.

Cronbach, L. J., and Azuma, H. Internal-consistency reliability formulas applied to randomly sampled single-factor tests: an empirical comparison. Educational and Psychological Measurement, 1962, 22, 645-665.

Cronbach, L. J., and Merwin, J. C. A model for studying the validity of multiple-choice items. Educational and Psychological Measurement, 1955, 15, 337-352.

Cronbach, L. J., and Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952, 17, 127-148.

Davis, M. Comparison of sequential bioassays in small samples. Journal of the Royal Statistical Society, Series B, 1971, 33, 78-87.

Dixon, W. J., and Mood, A. M. A method for obtaining and analyzing sensitivity data. Journal of the American Statistical Association, 1943, 43, 109-126.

Dupač, V., and Král, F. Robbins-Monro procedure with both variables subject to experimental error. The Annals of Mathematical Statistics, 1972, 43, 1089-1095.

Ferguson, R. L. Computer assistance for individualizing measurement. Technical Report. Pittsburgh, Pa.: University of Pittsburgh, 1971.

Fischer, G. H. Conditional maximum-likelihood estimation of item parameters for a linear logistic test-model. Research Bulletin No. 9. Vienna: Psychologisches Institut der Universität Wien, 1972.

Fischer, G. H., and Scheiblechner, H. H. Two simple methods for asymptotically unbiased estimation in Rasch's measurement model with two categories of answers. Research Bulletin No. 1. Vienna: Psychologisches Institut der Universität Wien, 1970.

Freeman, P. R. Optimal Bayesian sequential estimation of the median effective dose. Biometrika, 1970, 57, 79-89.

Guilford, J. P. Psychometric methods. (2nd ed.) New York: McGraw-Hill, 1954.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K., and Traub, R. E. Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.

Hansen, D. N., and Schwarz, G. An investigation of computer-based science testing. Tallahassee: Institute of Human Learning, Florida State University, 1968.

Hodges, J. L., Jr., and Lehmann, E. L. Two approximations to the Robbins-Monro Process. In J. Neyman (Ed.), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1. Berkeley: University of California Press, 1956.

Indow, T., and Samejima, F. LIS measurement scale for non-verbal reasoning ability. Tokyo: Nihon-Bunka Kagakusha, 1962. (In Japanese.)

Indow, T., and Samejima, F. On the results obtained by the absolute scaling model and the Lord model in the field of intelligence. Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University, 1966. (In English.)

Kesten, H. Accelerated stochastic approximation. Annals of Mathematical Statistics, 1958, 29, 41-59.

Kolakowski, D. Maximum likelihood estimation of item parameters and latent ability by generalized probit analysis. Paper presented at the spring meeting of the Psychometric Society, Princeton, N.J., 1969.

Kolakowski, D. An investigation of bias in the estimation of latent ability and item parameters under the normal ogive model. Paper presented at the meeting of the Psychometric Society, Princeton, N.J., March 1972.

Kolakowski, D., and Bock, R. D. A Fortran IV program for maximum likelihood item analysis and test scoring: Normal ogive model. Educational Statistics Laboratory Research Memo No. 12. Chicago: University of Chicago, 1970.

Lawley, D. N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, 273-287.

Lawley, D. N. The factorial analysis of multiple item tests. Proceedings of the Royal Society of Edinburgh, 1944, 62-A, 74-82.

Lazarsfeld, P. F. Latent structure analysis. In S. Koch (Ed.), Psychology: A study of a science. Vol. 3. New York: McGraw-Hill, 1959. Pp. 476-542.

Lees, D. M., Wingersky, M. S., and Lord, F. M. A computer program for estimating item characteristic curve parameters using Birnbaum's three-parameter logistic model. Office of Naval Research Technical Report, Contract No. N00014-69-C-0017. Princeton, N.J.: Educational Testing Service, 1972.

Levine, M. V. Transforming curves into curves with the same shape. Journal of Mathematical Psychology, 1972, 9, 1-16.

Linn, R. L., Rock, D. A., and Cleary, T. A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

Linn, R. L., Rock, D. A., and Cleary, T. A. Sequential testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

Lord, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F. M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75. (a)

Lord, F. M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548. (b)

Lord, F. M. Some perspectives on "The Attenuation Paradox in Test Theory." Psychological Bulletin, 1955, 52, 505-510.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50. (a)

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970. Pp. 139-183. (b)

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813. (b)

Lord, F. M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242. (c)

Lord, F. M. Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31. (d)

Lord, F. M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711. (e)

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Research Bulletin 72-00. Princeton, N.J.: Educational Testing Service, in press.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Meredith, W. Some results based on a general stochastic model for mental tests. Psychometrika, 1965, 30, 419-440.

Merwin, J. C. Rational and mathematical relationships of six scoring procedures applicable to three-choice items. Journal of Educational Psychology, 1959, 50, 153-161.

Mulaik, S. A. A mathematical investigation of some multidimensional Rasch models for psychological tests. Paper presented at the meeting of the Psychometric Society, Princeton, N.J., March, 1972.

Nishisato, S., and Torii, Y. Assessment of information loss in scoring monotone items. Multivariate Behavioral Research, 1971, 6, 91-103.

Odell, P. L. An empirical study of three stochastic approximation techniques applicable to sensitivity testing. NAVWEPS Report 7837. Albuquerque, N.M.: U.S. Naval Weapons Evaluation Facility, 1961.

Owen, R. J. Bayesian sequential design and analysis of dichotomous experiments with special reference to mental testing. Ann Arbor, Mich.: Author, 1970.

Panchapakesan, N. The simple logistic model and mental measurement. (Doctoral dissertation, University of Chicago) Chicago: University of Chicago Library, Dept. of Reproduction, 1969.

Paterson, J. J. An evaluation of the sequential method of psychological testing. Unpublished doctoral dissertation, Michigan State University, 1962.

Pollack, I. Methodological determination of the PEST (Parameter Estimation by Sequential Testing) procedure. Perception and Psychophysics, 1968, 3, 285-289.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogishe Institut, 1960.

Rasch, G. On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1961, 4, 321-333.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), Readings in mathematical social science. Chicago: Science Research Associates, 1966. Pp. 89-107. (a)

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57. (b)

Robbins, H., and Monro, S. A stochastic approximation method. The Annals of Mathematical Statistics, 1951, 22, 400-407.

Samejima, F. A general model for free-response data. Psychometric Monograph Supplement, 1972, No. 18.

Scheiblechner, H. H. CML-parameter-estimation in a generalized multi-factorial version of Rasch's probabilistic measurement-model with two categories of answers. Research Bulletin Nr. 4/71. Vienna: Psychologisches Institut der Universität Wien, 1971. (a)

Scheiblechner, H. H. A simple algorithm for CML-parameter-estimation in Rasch's probabilistic measurement model with two or more categories of answers. Research Bulletin Nr. 5/71. Vienna: Psychologisches Institut der Universität Wien, 1971. (b)

Seeley, L. C., Morton, M. A., and Anderson, A. A. Exploratory study of a sequential item test. Technical Research Note 129. Washington, D.C.: U. S. Army Personnel Research Office, 1962.

Shiba, S. Information transmission rate of psychological tests I. Japanese Journal of Psychology, 1969, 40, 68-75. (a)

Shiba, S. Information transmission rate of psychological tests II. Japanese Journal of Psychology, 1969, 40, 121-129. (b)

Shoemaker, D. M., and Osburn, H. G. A simulation model for achievement testing. Educational and Psychological Measurement, 1970, 30, 267-272.

Solomon, H. (Ed.) Studies in item analysis and prediction. Stanford: Stanford University Press, 1961.

Solomon, H. (Ed.) Item analysis, test design, and classification. Project No. 1327. Stanford, Calif.: Stanford University, U.S. Office of Education, Cooperative Research Program, 1965.

Stuckey, C. W., Hutton, C. L., and Campbell, R. A. Decision rules in threshold determination. Journal of the Acoustical Society of America, 1966, 40, 1174-1179.

Taylor, M. M., and Creelman, C. D. PEST: Efficient estimates on probability functions. Journal of the Acoustical Society of America, 1967, 41, 782-787.

Tinsley, H. E. A., and Dawis, R. V. A comparison of the Rasch item probability with three common item characteristics as criteria for item selection. Technical Report No. 3003. Minneapolis, Minn.: University of Minnesota, The Center for the Study of Organizational Performance and Human Effectiveness, 1972.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Tsutakawa, R. K. Block up-and-down method in bio-assay. Unpublished doctoral dissertation. Chicago, Ill.: University of Chicago, 1963.

Tsutakawa, R. K. Random walk design in bio-assay. Journal of the American Statistical Association, 1967, 62, 842-856. (a)

Tsutakawa, R. K. Asymptotic properties of the block up-and-down method in bio-assay. The Annals of Mathematical Statistics, 1967, 38, 1822-1828. (b)

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished doctoral dissertation. Lafayette, Ind.: Purdue University, 1970.

Wasan, M. T. Stochastic approximation. Cambridge: Cambridge University Press, 1969.

Waters, C. J. Preliminary evaluation of simulated branching tests. Technical Research Note 140. Washington, D.C.: U.S. Army Personnel Research Office, 1964.

Waters, C. W., and Bayroff, A. G. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

Wetherill, G. B. Sequential estimation of quantal response curves. Journal of the Royal Statistical Society, Series B, 1963, 25, 1-38.

Wetherill, G. B., Chen, H., and Vasudeva, R. B. Sequential estimation of quantal response curves: A new method of estimation. Biometrika, 1966, 53, 439-454.

Wetherill, G. B., and Levitt, H. Sequential estimation of points on a psychometric function. British Journal of Mathematical and Statistical Psychology, 1965, 18, 1-10.

Wood, R. The efficacy of tailored testing. Educational Research, 1969, 11, 219-222.

Wood, R., and Skurnik, L. S. Item banking. London: National Foundation for Educational Research in England and Wales, 1969.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968. Pp. 85-101.

Wright, B., and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.